1 OF 1

ED 083 991

DOCUMENT RESUME

ED 083 991 LI 004 530

AUTHOR Wainwright, Jane; Hills, Jacqueline
TITLE Book Selection from MARC Tapes.
INSTITUTION Association of Special Libraries and Information Bureaux, London (England). Research and Development Dept.
PUB DATE Feb 73
NOTE 48p.;(13 references)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Bibliographic Citations; *Information Retrieval; *Information Services; *Library Material Selection; On Line Systems; Relevance (Information Retrieval); Search Strategies
IDENTIFIERS Machine Readable Cataloging; MARC; SDI; *Selective Dissemination of Information

ABSTRACT

In April 1971 the Aslib Research and Development Department began a study on selective dissemination from MARC tapes. The aim of the project was to explore the technical and economic feasibility of providing selective notifications of current books, by extraction from MARC tapes, to specialized libraries. Typical of the potential customers envisaged would be Aslib member organizations. A comparison would be made of the utility of the various elements in the MARC record as search keys. The project was planned in six phases: (1) planning and program specification for MARC file creation and searching; (2) programing; (3) exploratory work with test file; (4) pilot operation with users; (5) analysis of results, conclusions, report; and, (6) market survey. This report covers the first five phases of the project. (Author/SJ)

ED 083991

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Jane Wainwright
Aslib
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

U S DEPARTMENT OF HEALTH EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

BOOK SELECTION FROM MARC TAPES

Jane Wainwright
Jacqueline Hills

Aslib Research and Development Department
February 1973

LI 004 530

FILMED FROM BEST AVAILABLE COPY

Contents

1. Introduction (Page 1)

2. Background and summary (Page 2)
   2.1 Other work on selective dissemination from MARC tapes
   2.2 Brief description of project.

3. Methodology (Page 6)
   3.1 Programs
   3.2 Exploratory work
   3.3 Pilot users
   3.4 Profile construction and testing
      3.4.1 General
      3.4.2 Natural language boolean profiles
      3.4.3 Further types of profile
   3.5 Running the profiles: assessments

4. Results (Page 16)
   4.1 Performance
      4.1.1 General
      4.1.2 Precision figures - BNB tape searches - LC tape searches
      4.1.3 Precision and recall figures - BNB tape searches - LC tape searches
      4.1.4 Precision and recall for various searches - BNB tape searches - LC tape searches
      4.1.5 General figures - Currency - Items scanned
   4.2 Times and costs for on-line system
      4.2.1 Weekly running costs
      4.2.2 Times for searching various fields
      4.2.3 Times for scanning BNB Weekly List
   4.3 Relations with a bureau

5. Use of batch system (Page 33)
   5.1 Other batch systems in the UK
   5.2 Costs of other systems
   5.3 MARCAS batch system costs
   5.4 Large systems
   5.5 Provision of standard interest profiles

6. Conclusions and recommendations for future work (Page 37)

6.1 Implications of findings about retrieval characteristics

6.1.1 Recall and precision

6.1.2 Currency

6.1.3 Profiling

6.2 Standard profiles

6.3 Survey of book selection methods and needs

6.4 Other recommendations for future work

6.4.1 Investigation of other systems for improved performance and availability

6.4.2 Pilot service to fifty users

6.4.3 Profile development and evaluation

References

Appendix I MARCAS User Manual (Not included in this copy but available on request)

Appendix II Potential for use with other MARC format tapes.

1. Introduction

In April 1971 the Aslib Research and Development Department began a study on selective dissemination from MARC tapes. The aim of the project was to explore the technical and economic feasibility of providing selective notifications of current books, by extraction from MARC tapes, to specialised libraries. Typical of the potential customers envisaged would be Aslib member organisations. A comparison would be made of the utility of the various elements in the MARC record as search keys.

The project was planned in six phases:

1 - planning and program specification for MARC file creation and searching

2 - programming

3 - exploratory work with test file

4 - pilot operation with users

5 - analysis of results, conclusions, report

6 - market survey

This report covers the first five phases of the project. The first three phases were financed by a research contract from the UK Office for Scientific and Technical Information, whose support is gratefully acknowledged.

Our thanks are due particularly to the libraries that acted as our 'users' - without their help and cooperation we would have been unable to carry out this study. Brian Skinner of Cybernet Time-Sharing Ltd and the staff of the British National Bibliography have given great help. Work on this project has been carried out by Jane Wainwright, Jackie Hills, Brian Vickery and, in the first phase, also by Suman Datta.

2. Background and Summary

2.1 Other work on selective dissemination from MARC tapes:

United Kingdom

During the summer of 1971 Aslib Research and Development Department conducted a survey of British National Bibliography (BNB) - MARC tape users in the UK.[1] From this survey it appeared that four organisations were carrying out selective dissemination from MARC tapes. The majority of organisations (group I in the list below) were using Dewey Decimal classification (DC) numbers or ranges of numbers for selection. One organisation (II) was using Library of Congress (LC) subject headings for retrieval. A few organisations (III) had plans for future work.

I Trinity College,Dublin - profiles for social scientists, Irish government departments and industrial information centres

Queens University,Belfast - a few group profiles

UKAEA,Aldermaston - a book pre-selection program,rather than SDI

II Birmingham Libraries Cooperative Mechanization Project - profiles for social scientists

III The City University,London
The Polytechnic of North London

In all of this work search of the tapes is in the 'batch' processing mode.

North America

The other work in this field is being carried out in the US and Canada. An early study was carried out at Indiana University by Studer[2], using MARC I records. The experimental SDI service, to forty social scientists,used weighted LC classification numbers and subject headings.

This was followed by an operational weekly MARC II SDI service at the Oklahoma Department of Libraries[3]. The search keys used are DC and LC classification numbers or ranges of numbers and the service is being used by over twenty groups in the US and Canada.

However since February 1971 Canada has had its own service. The Office of Technical Services Library, University of Saskatchewan, in cooperation with the National Science Library, National Research Council of Canada[4], has been carrying out selective dissemination of MARC - the SELDOM system. This is a highly flexible program, with seven main search keys (personal name, corporate name, title, DC and LC classification number, geographic code, date) offering boolean and weighted logic, and various output options

Mauerhoff[4] mentions work being done at Washington University School of Medicine on searching by LC classification numbers, and similar work at the University of Florida, Yale University Library, and Harvard University Library. From a survey of automated activities in US libraries[5] it appears that the University of Minnesota Libraries and the National Center for Atmospheric Research, Boulder, Colorado, are also studying this area.

The only report of on-line work comes from Syracuse University[6] where a research project was carried out using MARC I records with MOLDS, a generalized computer-based interactive retrieval program, which allowed for many different retrieval keys.

2.2 Brief description of project

If a service such as the one envisaged were run commercially it would be carried out in batch mode. However in this exploratory work we were considering various types

of search key and various forms of profile and would therefore be creating and modifying many profiles. The quickest and most effective way of doing this is to work in the on-line mode.

A specification for the requisite programs to handle MARC format tapes, on-line,with teletype-compatible terminal access, was prepared and submitted to a number of commercial computer time-sharing bureaux. Quotations were received from four bureaux: Cybernet Time Sharing Limited was selected on the basis of their relevant experience in on-line information retrieval work, and because they were able to amend existing software, thus offering the cheapest tender and the shortest completion time.

The MARC ASlib (MARCAS) programs allow both weighted and boolean searching on the Title and Author, LC classification number and subject headings, and the BNB Precis indexing terms and Reference Index Numbers, which uniquely identify each Precis term. The DC number as such is not used for searching, but a range of DC numbers may be employed to refine a search.

Exploratory work with a test file was begun on completion of the programming.

Twelve libraries, covering a range of subject fields, were contacted, and of these nine were used for the pilot operation. Profiles were constructed for these users and run on six weeks of BNB and six weeks of LC MARC tapes. The output from each weekly search was assessed for relevance by the users, and in the final week a measure of the recall of the system was also obtained.

The results of this pilot operation were then analysed,

in terms of precision and recall, for various combinations of the searchable fields. The best performance, with precision and recall both about 50%, was given by searching all verbal fields together - title and author, LC subject headings and Precis indexing terms (BNB tapes only).

Costs for the on-line system were identified.

A batch version of the MARCAS system was then implemented and computer costs of £1.30 per library, per week, for searching both BNB and LC tapes were calculated.

Further work which would be desirable background for implementing a cost recovery service include a survey of book selection and acquisition needs and methods currently used. Further study of available software, including testing, would probably indicate a more efficient system for operational use. The characteristics of standard profiles, their usefulness and relation to a specialized service also requires more study.

3. Methodology

3.1 Programs

An analysis of BNB-MARC tapes was carried out to gain information about various characteristics, such as the numbers of fields present, their average length, the numbers of records in certain subject fields, and the co-occurrence of indexing words in the title, DC and LC subject headings. Thought was also given to the fields that should be searchable; to the type of profile construction and the maximum number of terms that should be included; and to the subject areas that should be covered. This analysis led to the following decisions and implementations.

Initially an item is selected from a MARC tape for further processing only if it is in the English language, does not have a juvenile indicator, and does have a DC number. This last constraint was based on the fact that only low-level items are not assigned a DC number and, more particularly, that DC offers the most concise method of selecting subject areas. It was originally decided to include only items starting with DC numbers 0,3,5,6,or 7 since our interests were to be concentrated in the science, technology and social science fields. Items with DC number 9 were later included, in order to cover geography. (The only other possible field for subject area selection - the LC classification number - is not present on about 15 % of the items, and so its use was not explored). When the items have been selected from the original tape they are sorted into DC number order. Items assigned more than one DC number are duplicated in the file - once for each number (about 4% of the items are duplicates).

The following fields from each MARC record are stored on disc, and those marked with an asterisk* are searchable:

001 - International Standard Book Number (ISBN)
*050 - LC classification number
082 - DC number
*245 - Title and Author
260 - Place of publication, publisher and date
*650)- LC subject headings
651)
*690 - BNB Precis string
*692 - BNB Reference Index Number (RIN)

A range of DC numbers within which the search is carried out can also be specified.

A 'compressed' file is formed from the searchable elements in each record, in order to speed up the search process. The compressed file acts as a filter in that it eliminates non-hits, but each possible hit has to be checked against the main text file. Both files are linear. The fairly crude file structure is only feasible with small files. The search files are no longer held in the MARC format, which is basically a communications format. The file set-up procedure is carried out with two programs which are submitted for processing in remote job entry mode.

Searches can be made using either weighted or boolean logic on one or any combination of the searchable fields. Up to 30 terms can be used in a search. Searches can be entered directly from the keyboard or via paper tape or from a disc file. The output may contain the whole of each record that is stored, or any part of it that the searcher wishes, and may be typed by the terminal or put onto disc for printing in batch mode on the computer centre's line printer. Figure 1 indicates the available options.

Further modifications have been made to the original programs due both to our experience and Cybernet's continuing interest, and the system now provides a wide range of available options. The reader is referred to Appendix I - MARCAS User Guide - for full details of our current on-line retrieval system for MARC tapes.

Options for processing

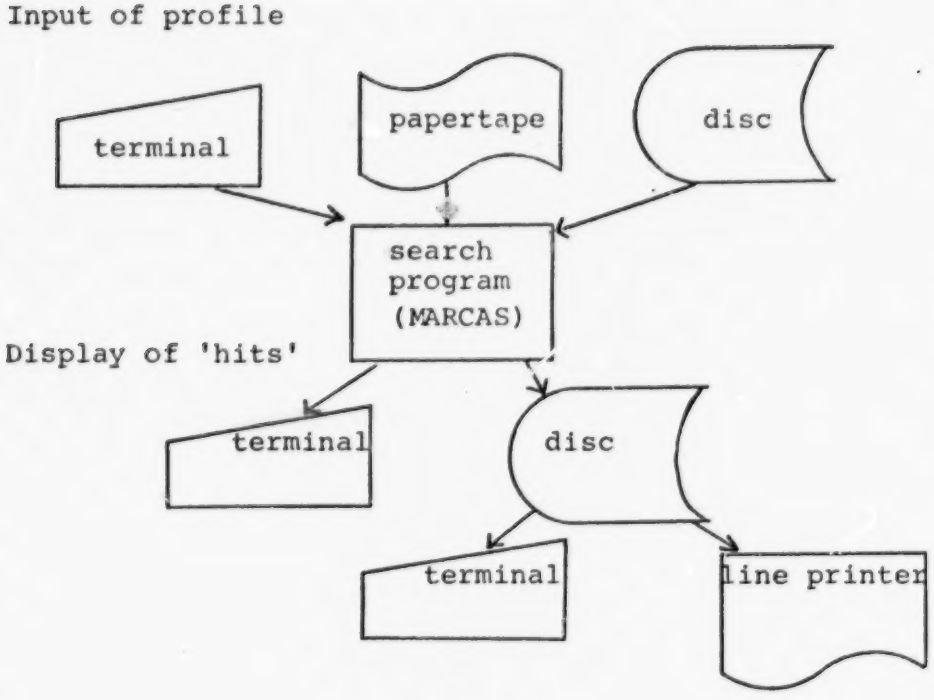

Figure 1

3.2 Exploratory work

By November 1971 the programs were ready for use, together with a test file, created from three weeks of BNB-MARC tapes, comprising some 1250 items. Some time was spent getting to know the system. It was obvious that the wide range of options offered by the programs would enable us easily to compare the use of different types of profile on the same base and also, because of the immediate response given by the system, that we would be able to test and amend profiles quickly and efficiently. The programs have proved fairly trouble-free for the whole duration of the project.

3.3 Pilot Users

Twelve libraries were originally approached and of these one did not reply, one proved too large to cope with and one, although keen at first, was unable to cooperate in the later stages of the project. Thus we had nine organisations (2 industrial/commercial, 2 nationalised industry, 1 local government, 2 research association/institute, 2 university, post-graduate).

The librarians were interviewed and a description of the main subject field of each library obtained. (These covered steel, plastics, electronics, instrumentation, computers, medicine, law, business and television.) We also gained an idea of the libraries' fringe interests - these varied from marketing, management and computing to library studies. We obtained an estimate of the number of book purchases per annum - this varied from 100 to 2,500, with an average of 650.

The librarians were asked about the sources they used for book selection. Only three of the nine libraries used the BNB Weekly List (one of these being the largest library). All libraries received and used publishers' announcements and catalogues, and most received specialised lists and scanned book reviews. Other sources mentioned included Bookseller, Times Literary Supplement, HMSO List and

Aslib Book List. The other major source is, naturally, requests from readers. The librarians who used BNB foundit 'a bit late but useful nevertheless.' Most libraries had some standing orders for series, etc.

We also obtained a list of recent book acquisitions.

The librarians or their assistants marked the three BNB Weekly Lists corresponding to our test-file at two levels of relevance (1 - books they would buy, 2 - books they were interested in knowing about).

3.4 Profile construction and testing

3.4.1 General

The profile construction aids which we had available were:

- the 18th edition of the Dewey Decimal Classification
- the Library of Congress Classification
- card file of terms used in the 1971 Precis index, with their related RIN's
- recent acquisitions lists from our test libraries
- three week BNB-MARC test file (held on-line)

The original plan had been to construct profiles from each library's described interests, using the first four aids above and then to test these profiles against the MARC test file. However it proved difficult to construct profiles in this way because, on the whole, not enough information was available. In fact the known relevant items in the test file provided us with the best guide

for constructing profiles*, the other aids being used mainly for producing synonyms and related terms or concepts.

Profiles were constructed in 'natural language' rather than in specific terms from DC or LC classifications since we required to test all the verbal fields for retrieval. We also felt that we could handle with more confidence the broader class formulations. In fact we found the LC classification fairly difficult to deal with - as did another researcher[7] who used only the LC subject headings for profiling.

3.4.2 Natural language boolean profiles

Eventually, though this stage proved more time-consuming than anticipated, natural language boolean profiles were constructed for all libraries. An example is shown in Figure 2:

'BIOCHEMIST' OR 'BIOCHEMICAL' OR 'BIOLOG' AND ('CHEMIST' OR 'DEVELOP' OR 'EXPERIMENT' OR 'PHYSIC' OR 'CONTROL') AND NOT 'SCHOOL' OR 'BIOPHYSIC' OR 'BIOMEDICAL ENGINEER' OR 'ERGONOMIC' OR 'DESIGN ANTHROPOMETRY' OR 'MEDICAL' AND ('TECHNOLOG' OR 'LABORATOR' OR 'EQUIPMENT') AND NOT 'PROFESSION' :

Figure 2

Since each profile that can be entered in the MARCAS system has a limit of 30 terms, several profiles had to be constructed for each library (the number varied from two to six, with an average of four). In fact the average number of terms needed to describe a library's interests was 100 (with a range from 50 to 150).

*UKCIS have also found that the examination of known relevant items has proved the most useful aid in profile construction.

The initial profile creation took about five days for each library. These profiles were then tested on-line against the three week test file (searching on the Title and Author field (245), the LC subject headings (650 and 651) and the BNB Precis string (690)) and amended and refined as necessary. This latter task was achieved very quickly and painlessly with the on-line facilities. The profiles were entered using paper tape.

3.4.3 Further types of profile

For five of the libraries further profiles were constructed. An example of each type is illustrated below.

i) Natural language, weighted - for searching the 245, 650/1, and 690 fields

```
'TELECOMMUNICAT' (6),'TELEVISION'(6),' TV '(6),
' T.V. '(6),'RADIO '(6),'RADIOCOMMUNICAT'(6),
'WIRELESS'(6),'BROADCAST'(6),'TRANSMITTER'(6),
'RECEIVER'(6),'RECEIVING SET'(6),'AERIALS'(6),
'ANTENNAS'(6),'THYRISTORS'(6),'ELECTRONIC EQUIPMENT'(6)
'PRINTED CIRCUIT'(6),'STEEL'(3),'STRUCTUR'(3),
'METRIC SYSTEM'(4),'BUILDING'(2),'POST OFFICE'(4),
'GREAT BRITAIN'(1),'ANNUAL REPORT','SCHEDULING'(6);
```

Figure 3

ii) LC classification numbers, weighted - for searching the 050 field

```
'LC5201','LC5209','LC5215','LC5219','LC5256','LC657* '
'LC6581','PN1991','PN1992','PN1993','PN1994','PN47** '
'PN511* ','PN512* ','PN513* ','PN514* ','PN5150'
;
```

Figure 4

iii) Reference Index Numbers, weighted - for searching the 692 field

```
'000204072','000706256','00380402X','002002043',
'003005062','001106252','001401157','001901044',
'000102083','00010325X','000303127','001702068',
'001805177','001006045','001001205','001007041',
'003106209','002208016','00200318X','000204056',
'000106038'(-5);
```

Figure 5

These profiles were also tested against the three week test file and altered where required.

The three types of profile above, together with the natural language boolean one (3.4.2), enabled us to explore all our chosen searchable fields and also to compare the usefulness of boolean logic with weighted logic.

3.5 Running the profiles: assessments

The natural language boolean profiles for each library were run on six weeks of BNB tapes and six weeks of LC tapes, with no further amendments to the profiles. This work was done on-line (with output onto a disc for computer line-printer output) for our convenience, although, if an operational system had existed the task could have been carried out equally well in batch mode. Central processor and connection times for the searches were recorded for two weeks.

The results of the first five weeks' searches (an example of the output is shown in Figure 6) were sent to the libraries for assessment at three levels of relevance:

1 - Important reference; books you would buy
2 - Relevant reference; books you are interested in knowing about
3 - Non relevant

The users were also asked to mark those items they already knew of, so that we might get an idea of the currency of the tapes.

For the final week the users were sent the complete BNB Weekly List and a copy of the LC file, and asked to mark relevant items either 1 or 2 as above. This enabled us to obtain a value for recall. Unfortunately one library did not have the time to mark up the last three weeks' LC files.

The other types of profile (3.4.3) were run only on the last week of the BNB and LC tapes. The users were not sent the output from these searches since we already had their assessment of the total relevant items in these files.

```
***   1  72175103
      82   611  .0182
     245   THE HUMAN ADIPOSE CELL  A MODEL FOR ERRORS IN METABOLIC REGULATION  IBY- DAVID J. GALTON
     260   NEW YORK  APPLETON-CENTURY-CROFTS  IC1971-
     650   ADIPOSE TISSUES
     650   CORPULENCE
     650   METABOLISM, INBORN ERRORS OF
***   1  73126889
      82   611  .0184
     245   BONE MARROW AND BONE TISSUE  COLOR ATLAS OF CLINICAL HISTOPATHOLOGY  FOREWORD BY W. STICH. TRANSLATED BY H.
           J. HIRSCH
     260   BERLIN  NEW YORK  SPRINGER-VERLAG  1971
     650   DIAGNOSIS, CYTOLOGIC
     650   MARROW  ATLASES
***   1  73172082
      82   611  .0189  31
     245   ORBAN'S ORAL HISTOLOGY AND EMBRYOLOGY  EDITED BY HARRY SICHER IAND- S. N. BHASKAR
     260   SAINT LOUIS  MOSBY  1972
     650   MOUTH
     650   TEETH
     650   JAWS
     650   SALIVARY GLANDS
***   1  74883704
      82   612  .8  042
     245   SECOND INTERNATIONAL MEETING OF THE INTERNATIONAL SOCIETY FOR NEUROCHEMISTRY  MILAN, SEPTEMBER 1-5, 1969
           EDITED BY R. PAOLETTI, R. FUMAGALLI, IAND- C. GALLI
     260   MILANO  TAMBURINI  [1969-
     650   NEUROCHEMISTRY  CONGRESSES
***   1  75835866
      82   615  .65
     245   SEROLOGICAL AND IMMUNOLOGICAL METHODS  TECHNICAL MANUAL OF THE CANADIAN RED CROSS BLOOD TRANSFUSION SERVICE
           BY B. P. L. MOORE, PATRICIA HUMPHREYS IAND- CECILE A. LOVETT-MOSELEY
     260   TORONTO  CANADIAN RED CROSS SOCIETY  1968, REPRINTED 1969
     650   BLOOD  ANALYSIS AND CHEMISTRY
     650   BLOOD  TRANSFUSION
     650   IMMUNOHEMATOLOGY
***   1  79178772
      82   546
     245   SYNTHESES AND PHYSICAL STUDIES OF INORGANIC COMPOUNDS  BY C. F. BELL
     260   OXFORD  NEW YORK  PERGAMON PRESS  [1972-
     650   CHEMISTRY, INORGANIC
***   1  73884461
      82   553  .19
     245   PROBLEMS OF HYDROTHERMAL ORE DEPOSITION  THE ORIGIN, EVOLUTION AND CONTROL OF ORE-FORMING FLUIDS  SYMPOSIUM
           ORGANIZED BY THE INTERNATIONAL ASSOCIATION ON THE GENESIS OF ORE DEPOSITS, ST. ANDREWS, SCOTLAND, 1967. EDITED
           BY ZDENEK POUBA AND MIROSLAV STEMPROK
     260   STUTTGART  E. SCHWEIZERBART  1970
     650   HYDROTHERMAL DEPOSITS  CONGRESSES
***   1  76161890
      82   575.1
     245   HEREDITY AND DEVELOPMENT  IBY- JOHN A. MOORE
     [illegible]
```

4. Results

4.1 Performance

4.1.1 General

The measures used to evaluate the performance of the system are recall and precision (expressed as percentages). The figures for each separate library have been taken as the ratio of the total values over the six weeks (on the assumption that the results are reasonably homogeneous over the six weeks) (section 4.1.2). For the figures in sections 4.1.3 and 4.1.4 only the final week's values are concerned. The overall figures for all libraries have been taken as the average of the ratios of each library (because each library's figures should be allowed to have an equal effect on the final figures and in fact some libraries had large outputs and some small). Rather than being left out of the calculations recall and precision ratios of $\frac{0}{0}$ have been taken as 100%, since the relationship between the first and second levels of relevance then makes more sense. In Tables 1 - 4 the actual number of relevant retrieved items has been noted.

The results fall into three main groups. The first covers precision figures for all libraries over six weeks on natural language boolean searches, the second precision and recall figures for all libraries for the final week on natural language boolean searches, and the third precision and recall figures for five libraries for the final week for various types of search.

4.1.2 Precision figures (Table 1)

Table 1 shows precision figures for natural language boolean searches for each library over six weeks on BNB and LC tapes. The precision figures are given at two levels of relevance -

P1 - is the ratio (expressed as a percentage) of the relevant retrieved items at relevance level 1 (see section 3.5) to the total retrieved items.

P2 - is the ratio of the relevant retrieved items at relevance level 1 plus 2 to the total retrieved items.

Average values over all libraries have been calculated.

BNB tape searches

The fields searched were the Title and Author (245), the LC subject headings (650 and 651) and the Precis indexing terms (690). Precision figures for searching various combinations of these fields h ve been identified. If we consider the precision averages P2 there is an increase in precision on searching the 650/1 field as opposed to the 245 field -from 59.1% (245) to 65.8% (650/1). When all three 'verbal' fields (i.e. including the 690) are searched the precision drops a little (to 55.7%) but this is probably compensated by an increase in recall (see 4.1.3). Overall recall figures are not available because the task of obtaining these figures would have put too great a burden on the users. However the relative recall (RR2) of the 245 and 650/1 fields to the total number of relevant items retrieved by searching all three fields (at relevance levels 1 plus 2) is seen as 55.3% and 62.1% respectively.

LC tape searches

The Precis terms are not present on the American tapes so we can only compare searches on the 245 and 650/1 fields. It is interesting to note that although the average precision value P1 is fairly low (about 10%) the value for P2 is nearly 50%. In other words though libraries may not wish to buy many of these American books, they are glad to have information about them. (see also 4.1.5 - Currency).

Table 1. Natural language boolean searches over six weeks.

Library	Precision								Relative Recall	
	Fields searched									
	Title LC Subject headings Precis*		Title		LC Subject headings		Title LC Subject headings		Title	LC Subject headings
BNB Tape	P1	P2	P1	P2	P1	P2	P1	P2	RR2	RR2
1	8.5	84.5	10.8	83.8	9.5	90.5	8.5	84.7	51.7	63.3
2	16.4	70.9	37.5	75.0	31.8	77.3	25.0	75.0	46.2	43.6
3	11.5	15.4	25.0	33.0	30.0	30.0	15.8	21.1	100.0	75.0
4	19.6	47.8	20.8	50.0	29.4	64.7	20.0	56.0	54.5	50.0
5	28.3	63.0	43.3	83.3	43.3	80.0	31.7	65.9	86.2	82.8
6	58.7	71.6	74.8	86.9	69.1	84.6	70.5	85.5	49.2	60.8
7	4.8	20.6	5.3	15.8	8.6	28.6	5.4	21.4	46.2	76.9
8	72.5	91.3	50.0	66.7	82.8	93.1	72.7	87.9	6.3	42.9
9	8.1	36.0	10.8	37.8	9.9	43.7	7.9	36.8	57.1	63.3
	218.4	501.1	278.3	532.3	314.4	592.5	257.5	534.3	497.4	558.6
Average	25.4	55.7	30.9	59.1	34.9	65.8	28.6	59.4	55.3	62.1
Total Relevant Retrieved	259	468	127	221	160	276	187	344	221	276
LC Tape	P1	P2	P1	P2	P1	P2			RR2	RR2
1	0	81.9	0	82.1	0	92.6			60.4	82.4
2	17.0	56.6	26.7	66.7	25.0	62.5			66.7	66.7
3	7.7	23.1	10.0	30.0	14.3	42.9			100.0	100.0
4	10.8	39.2	12.9	40.3	13.7	41.2			86.2	72.4
5	20.4	50.0	30.0	60.0	27.8	52.8			66.7	70.4
6	1.9	31.7	0	30.1	2.5	36.4			43.1	86.3
7	3.7	13.8	5.7	20.0	4.5	15.7			46.7	93.3
8	18.9	58.9	20.0	60.0	18.4	69.4			48.2	60.7
9+	0	31.1	0	42.9	0	32.6			63.2	78.9
	80.4	386.3	105.3	432.1	106.2	446.1			581.2	711.1
Average	8.9	42.9	11.7	48.0	11.8	49.6			64.6	79.0
Total Relevant Retrieved	54	321	37	189	42	245			189	245

4.1.3 Precision and recall figures (Table 2)

Table 2 shows precision and recall figures for natural language boolean searches for each library for the final week only, on BNB and LC tapes. The figures are at two levels: P1, R1 (precision and recall at relevance level 1) and P2, R2 (precision and recall at relevance levels 1 plus 2). Average values over all libraries have been calculated.

BNB tape searches

Searches over all verbal fields at relevance levels 1 plus 2 give 42.0% precision and 56.3% recall. Searches on the 245 field alone improved precision but lowered recall (44.8% and 34.3% respectively), while these on the 650/1 field improved precision even more and did not lower recall so much (50.8% and 42.1% respectively). Recall is obviously being increased (R1 in Table 2) by the use of the 690 field - in fact by 42.1% at relevance level 1 and 24.2% at relevance levels 1 plus 2.

LC tape searches

Precision and recall figures for searches on the LC tapes show the same trends as those on BNB tapes. Searching both the 245 and the 650/1 fields gives precision and recall of 46.3% and 59.7% respectively; searching the 245 field alone gives 51.3% and 44.8% respectively; searching the 650/1 field gives 49.8% and 44.7% respectively. These are the figures for relevance levels 1 plus 2 - those for relevance level 1 are much lower.

Table 2. Natural language boolean searches for final week.

Library	Precision and Recall												Recall increase using Precis field	
	Fields searched													
	Title, LC Subject headings, Precis*				Title				LC Subject headings					
BNB Tape	P1	R1	P2	R2	P1	R1	P2	R2	P1	R1	P2	R2	RI1	RI2
1	0	0/0	61.5	80.0	0	0/0	57.1	40.0	0	0/0	71.4	50.0	0/0	30.0
2	0	0	60.0	42.4	0	0	0	0	0	0	33.3	14.3	0	28.6
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	20.0	100.0	30.0	100.0	66.7	100.0	66.7	66.7	50.0	100.0	75.0	100.0	0	0
5	9.1	100.0	81.8	75.0	9.1	100.0	81.8	75.0	10.0	100.0	80.0	66.7	0	0
6	45.7	24.6	71.4	25.5	63.6	10.8	90.9	10.2	37.5	9.2	87.5	14.3	12.3	9.2
7	0	0/0	0	0/0	0	0/0	0	0/0	0	0/0	0	0/0	0/0	0/0
8	66.7	66.7	66.7	50.0	0/0	0	0/0	0	0/0	0	0/0	0	66.7	50.0
9	0	0/0	6.9	33.3	0	0/0	7.1	16.7	0	0/0	10.0	33.3	0/0	0
	141.5	591.3	378.3	506.7	239.4	510.8	403.6	308.6	197.5	509.2	457.2	378.6	379.0	217.8
Average	15.7	65.7	42.0	56.3	26.6	56.8	44.8	34.3	21.9	56.6	50.8	42.1	42.1	24.2
Total Rel.Ret	21	21	52	52	10	10	25	25	9	9	33	33	10	16
LC Tape	P1	R1	P2	R2	P1	R1	P2	R2	P1	R1	P2	R2		
1	0	0/0	55.6	71.4	0	0/0	70.0	50.0	0	0/0	63.6	50.0		
2	15.4	100.0	38.5	83.3	25.0	100.0	50.0	66.7	16.7	50.0	33.3	33.3		
3	0	0/0	100.0	00.0	0	0/0	100.0	100.0	0	0/0	100.0	100.0		
4	0	0	35.0	70.0	0	0	38.9	70.0	0	0	33.3	50.0		
5	33.3	66.7	50.0	30.0	37.5	50.0	62.5	25.0	37.5	50.0	50.0	20.0		
6	7.7	40.0	34.6	36.0	0	0	22.2	8.0	10.0	40.0	45.0	36.0		
7	0	0	13.0	50.0	0	0	16.7	16.7	0	0	27.3	50.0		
8	30.4	53.8	43.5	37.0	41.7	38.5	50.0	22.2	27.3	23.1	45.5	18.5		
9+														
	86.8	460.5	370.2	477.7	104.2	388.5	410.3	358.6	91.5	363.1	398.0	357.8		
Average	10.9	57.6	46.3	59.7	13.0	48.6	51.3	44.8	11.4	45.4	49.8	44.7		
Total Rel.Ret.	15	15	51	51	10	10	33	33	9	9	36	36		

*Field not present on LC tapes

+ Assessment not given

4.1.4 Precision and recall for various searches

BNB tape searches (Table 3)

Precision and recall figures (P2, R2) for relevance levels 1 plus 2, are given for natural language boolean and weighted searches, for LC classification number (field 050) searches (using weighted logic) and for Reference Index Number (RIN) (field 692) searches (weighted logic). These searches were carried out on the final week's tape and for only five libraries; average values over these five are calculated.

For the natural language searches precision and recall values are given for all the various combinations of searching the 245, 650/1 and 690 fields. For both boolean and weighted logics the combination of 650/1 and 690 searching produced the best results, while searching on the 245 field alone gave the worst performance. Boolean logic performed slightly better than weighted logic. Searches on the LC classification numbers gave an increase in precision, but a decrease in recall. Searches on the RIN produced lower precision and lower recall.

LC tape searches (Table 4)

Precision and recall figures are given for natural language boolean and weighted searches and for LC classification number searches. These searches were carried out on the final week's tape and for five libraries; one library failed to return the relevance assessment so the average values are calculated over four libraries.

The values here are much higher than those for similar BNB tape searches. This is partly due to library 3, which had no relevant items on the BNB tape and only one (which was retrieved) on the LC tape, and also to library 5, which had low values for the BNB tape and gave no figures for the LC tape.

Table 3. Various searches for BNB final week.

Library	Precision and Recall													
	Fields searched													
	Title LC Subject headings Precis		Title		LC Subject headings		Precis		Title LC Subject headings		Title Precis		LC Subject headings Precis	
NATURAL LANGUAGE BOOLEAN SEARCH	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2
1	61.5	80.0	57.1	40.0	71.4	50.0	70.0	70.0	50.0	50.0	58.3	70.0	12.7	80.0
2	60.0	42.9	0	0	33.3	14.3	60.0	42.9	33.3	14.3	60.0	42.9	60.0	42.9
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	30.0	100.0	66.7	66.7	75.0	100.0	22.2	66.7	75.0	100.0	22.2	66.7	30.0	100.0
5	6.9	33.3	7.1	16.7	10.0	33.3	14.3	33.3	7.7	33.3	9.5	33.3	8.3	33.3
	158.4	256.2	130.9	123.4	189.7	197.6	166.5	212.9	166.0	197.6	150.0	212.9	171.0	256.2
Average	31.7	51.2	26.2	24.7	38.0	39.5	33.3	42.6	33.2	39.5	30.0	42.6	34.2	51.2
Total rel.ret.	16	16	7	7	11	11	14	14	11	11	14	14	16	16
NATURAL LANGUAGE WEIGHTED SEARCH	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2
1	60.0	60.0	50.0	40.0	100.0	40.0	100.0	70.0	50.0	50.0	58.3	70.0	75.0	90.0
2	40.0	28.6	0	0	33.3	14.3	50.0	28.6	25.0	14.3	40.0	28.6	50.0	28.6
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	30.0	100.0	22.2	66.7	30.0	100.0	22.2	66.7	30.0	100.0	22.2	66.7	30.0	100.0
5	7.4	33.3	0	0	15.4	33.3	13.3	33.3	10.5	33.3	8.7	33.3	10.5	33.3
	137.4	221.9	72.2	106.7	178.7	187.6	185.5	198.6	115.5	197.6	129.2	198.6	165.5	251.9
Average	27.5	44.4	14.4	21.3	35.7	37.5	37.1	39.7	23.1	39.5	25.8	39.7	33.1	50.4
Total Rel.Ret.	13	13	6	6	10	10	13	13	11	11	13	13	16	16

	LC Classif. No.	
LC CLASS. NO.WEIGHT	P2	R2
1	66.7	66.7
2	66.7	28.6
3	0	0

	RIN	
RIN WEIGHTED SEARCH	P2	R2
1	54.5	60.0
2	50.0	42.9
3	0	0

Library	Precision and Recall													
	Fields searched													
	Title LC Subject headings Precis		Title		LC Subject headings		Precis		Title LC Subject headings		Title Precis		LC Subject headings Precis	
NATURAL LANGUAGE BOOLEAN SEARCH	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2
1	61.5	80.0	57.1	40.0	71.4	50.0	70.0	70.0	50.0	50.0	58.3	70.0	12.7	80.0
2	60.0	42.9	0	0	33.3	14.3	60.0	42.9	33.3	14.3	60.0	42.9	60.0	42.9
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	30.0	100.0	66.7	66.7	75.0	100.0	22.2	66.7	75.0	100.0	22.2	66.7	30.0	100.0
5	6.9	33.3	7.1	16.7	10.0	33.3	14.3	33.3	7.7	33.3	9.5	33.3	8.3	33.3
	158.4	256.2	130.9	123.4	189.7	197.6	166.5	212.9	166.0	197.6	150.0	212.9	171.0	256.2
Average	31.7	51.2	26.2	24.7	38.0	39.5	33.3	42.6	33.2	39.5	30.0	42.6	34.2	51.2
Total rel ret.	16	16	7	7	11	11	14	14	11	11	14	14	16	16
NATURAL LANGUAGE WEIGHTED SEARCH	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2	P2	R2
1	60.0	60.0	50.0	40.0	100.0	40.0	100.0	70.0	50.0	50.0	58.3	70.0	75.0	90.0
2	40.0	28.6	0	0	33.3	14.3	50.0	28.6	25.0	14.3	40.0	28.6	50.0	28.6
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	30.0	100.0	22.2	66.7	30.0	100.0	22.2	66.7	30.0	100.0	22.2	66.7	30.0	100.0
5	7.4	33.3	0	0	15.4	33.3	13.3	33.3	10.5	33.3	8.7	33.3	10.5	33.3
	137.4	221.9	72.2	106.7	178.7	187.6	185.5	198.6	115.5	197.6	129.2	198.6	165.5	251.9
Average	27.5	44.4	14.4	21.3	35.7	37.5	37.1	39.7	23.1	39.5	25.8	39.7	33.1	50.4
Total Rel. Ret.	13	13	6	6	10	10	13	13	11	11	13	13	16	16

	LC Classif. No.	
IC CLASS. NO. WEIGHT	P2	R2
1	66.7	66.7
2	66.7	28.6
3	0	0
4	66.7	66.7
5	14.3	33.3
	214.4	195.3
Average Total Rel. Ret.	42.9	39.1
	12	12

	RIN	
RIN WEIGHTED SEARCH	P2	R2
1	54.5	60.0
2	50.0	42.9
3	0	0
4	33.3	33.3
5	15.4	33.3
	153.2	169.5
Average Total Rel. Ret.	30.6	33.9
	12	12

Table 4. Various searches for LC final week.

Library	Precision and Recall					
	Fields searched					
	Title, LC Subject headings		Title		LC Subject headings	
NATURAL LANGUAGE BOOLEAN SEARCH	P2	R2	P2	R2	P2	R2
1	55.6	71.4	70.0	50.0	63.6	50.0
2	38.5	83.3	50.0	66.7	33.3	33.3
3	100.0	100.0	100.0	100.0	100.0	100.0
4	35.0	70.0	38.9	70.0	33.3	50.0
5+						
	229.1	324.7	258.9	286.7	230.2	233.0
Average	57.3	81.2	64.7	71.7	57.6	58.3
Total Relevant Retrieved	23	23	19	19	15	15
NATURAL LANGUAGE WEIGHTED SEARCH	P2	R2	P2	R2	P2	R2
1	71.4	71.4	70.0	50.0	70.0	50.0
2	45.5	83.3	66.7	66.7	33.3	33.3
3	16.7	100.0	50.0	100.0	33.3	100.0
4	34.6	90.0	34.6	90.0	33.3	70.0
5+						
	168.2	344.7	221.3	306.7	169.9	253.3
Average	42.1	86.2	55.3	76.7	42.5	63.3
Total Relevant Retrieved	25	25	21	21	17	17
	LC Classif. No.					
LC CLASSIFICATION NO. WEIGHTED SEARCH	P2	R2				
1	83.3	35.7				
2	40.0	33.3				
3	50.0	100.0				
4	58.3	70.0				
5+						
	231.6	239.0				
Average	57.9	59.8				
Total Relevant Retrieved	15	15				

+ Assessment not given

However it can be seen that searching the 245 and 650/1 fields in combination gave the best performance for natural language searches, with boolean searching slightly ahead of wieghted searching. Again LC classification number searches increased precision, though only marginally, but decreased recall.

Venn diagrams (Figure 7)

Another way of considering the results is by the use of Venn diagrams. The results of searching the final week's BNB and LC tapes by natural language boolean profiles are shown for four libraries. All the retrieved items have been entered in the appropriate relevant (relevance level 1 plus 2) or non-relevant diagrams. The sample used is rather small but it is interesting to note that in the BNB searches 86% of the relevant items could have been retrieved by searching only the 690 field and the additional 14% by searching the 650 field. The 690 field search also produces 85% of the non-relevant items, but since the precision is fairly high this should not be considered much of a disadvantage. For the LC searches however it appears that searching the title (245 field) will give 82% of the relevant items (and 66% of the non-relevant items).

4.1.5 General figures (Table 5)

Currency

Our attempt to obtain an idea of the currency of the service should not be considered too strictly as we were not running the tapes immediately after

Figure 7. Natural language boolean searches for final week.

Totals for libraries 1,2,3,4.

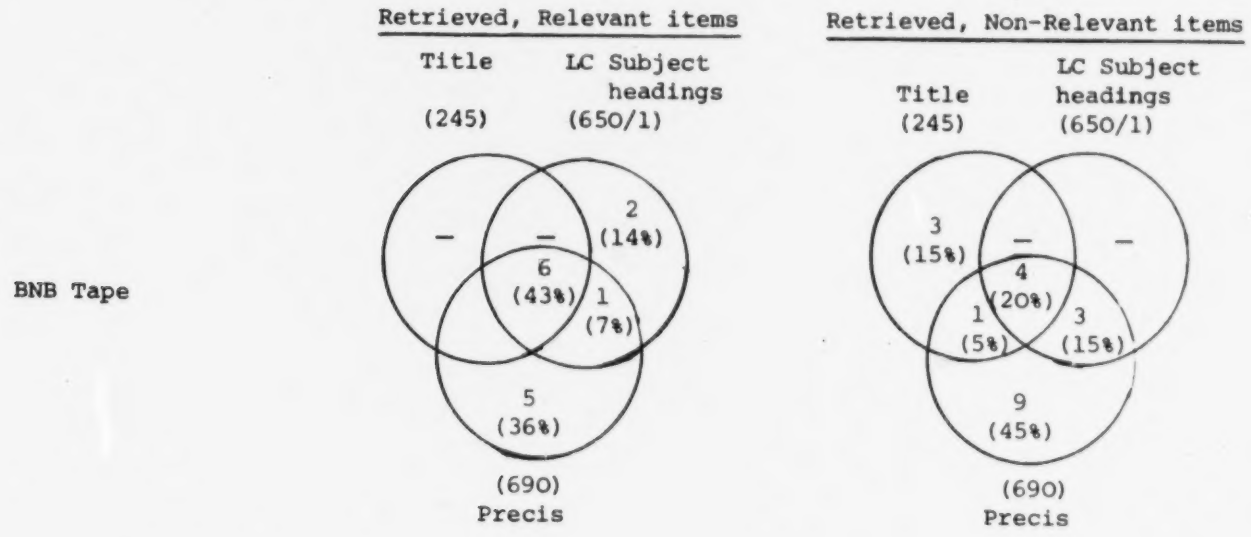

Table 5. General figures.

Library	Currency			Items scanned
	% already seen at each relevance level			Per week ratio: items scanned/items retrieved/items retrieved and relevant
BNB Tapes	I	R	N	
1	16.7	3.7	0	300/12/10
2	88.9	6.7	0	251/9/7
3	66.7	0	0	251/4/1
4	22.2	0	0	142/8/4
5	92.3	6.3	0	300/8/5
6	89.7	41.2	50.7	411/44/32
7	100	30.0	0	155/11/2
8	96.0	53.8	0	158/12/11
9	45.5	2.6	0	300/23/8
	618.0	144.3		2268/131/80
Average	68.7	16.0		252/15/9
LC Tapes	I	R	N	Items Scanned
1	$\frac{0}{0}$*	4.4	5.0	496/19/15
2	0	4.8	0	433/9/5
3	100	0	0	433/2/1
4	12.5	0	0	188/12/5
5	27.3	0	0	496/9/5
6	33.3	0	0	676/27/9
7	25.0	0	0	219/18/3
8	33.3	0	0	308/16/9
9	$\frac{0}{0}$	0	0	496/20/3
	231.1	9.2		3745/132/55
Average	33.0	1.0		416/15/6

* $\frac{0}{0}$ Figures have been left out of the average.

their production so, at least, those libraries receiving the BNB Weekly List should already have noted relevant items. However the averages (taken over five weeks) indicate that although libraries had already seen mention of 68.7% of the British books that they would in fact buy, they had only noted 16.0% of those they were interested in knowing about. (One library - number 6 - had already seen 50% of their non-relevant books!). As suspected fewer of the American books were previously known - 33.0% of important books and only 1% of interesting books. This latter is particularly remarkable since the precision value of these books is quite high.

Items scanned

We also considered the number of items that a library would have to look at in the book selection process. If we assume that a librarian will look through the whole of each Dewey Decimal class that he is interested in from the BNB Weekly List (this is probably not strictly true but is the only possible assumption that one can make here) then we can find the average number of items he would have to scan each week. We can also calculate the average number of items that are retrieved for one library each week and which of these are relevant. We arrive at an average ratio per week for items scanned: items retrieved: items retrieved and relevant of 252:15:9 for British books and 416:15:6 for American books. (The averages are taken over six weeks.)

4.2 Times and costs for on-line system

Our system has so far mainly been working in the on-line mode and any proposed commercial service would be run in batch mode. Thus costs for this system, although interesting, should not be taken as an indication of the costs for a service. For estimates of batch processing costs the reader should refer to section 5.

4.2.1 Weekly running costs

For the first two weeks' runs on both BNB and LC tapes of the natural language boolean searches, the computer usage time was noted (Table 6). This time is composed of central processor unit (CPU) time and connect or log-on time. For BNB tapes the average value for one library for one week was 42 seconds CPU time and 21 minutes connect time; for LC tapes it was 35 seconds CPU time and 18 minutes connect time (these figures may be lower than average as the first two LC tapes had fewer items than later ones). If we used Cybernet's most expensive rates (we were paying considerably less in fact) of £4.80 per minute CPU and £2.70 per hour connect we would have an average cost of about £4 per week for each library for each tape. The cost we were paying was in fact about £1.70 per library per tape.

However, a new file structure was implemented towards the end of the project, and although all the processing had been carried out by this time, a few tests were run on these new files (see last column, Table 6). These figures indicate that the CPU time for searching would be halved at least. This would give an average cost of £2.60 per week for each library, for each tape, run on-line (at the highest charge rate).

Table 6. Weekly running costs.

Library	Computer time		Final Week
	CPU (secs)	Connect (mins)	New files
BNB Tapes			CPU (Secs)
1	56	28	24
2	40	20	9
3	41	17	11
4	15	6	9
5	46	20	14
6	82	42	31
7	23	11	12
8	19	14	4
9	54	30	28
	376	188	142
Average	42	21	16
LC Tapes	CPU	Connect	CPU
1	46	25	26
2	33	15	15
3	30	15	13
4	16	7	11
5	38	15	19
6	62	30	35
7	24	14	16
8	18	12	13
9	50	25	31
	317	158	179
Average	35	18	20

The other weekly cost that can be easily identified is the cost of setting up the files, the work being carried out partly on-line and partly in remote job entry mode. This worked out at about £20 for each tape (again using the highest rates), or £15 at the rates we were paying. (The new file structure mentioned above would not alter this figure.)

4.2.2 Times for searching various fields

Though work has been done on the performance efficiency of searching the various fields available, not much effort has been devoted to considering the relative speeds of searching these fields because there are so many variables in the system. For a discussion of these variables in relation to the Scisearch system run by Cybernet (of which our system is an adaption) the reader should consult a report by Datta and Robertson[8]. The main factors affecting the search times in our system are:

i) number of terms, length of terms and logic used in profile
ii) number of items searched
iii) number of fields searched
iv) number of hits
v) number of fields printed out
vi) usage of disc input or output
vii) time of day that search was conducted (the computer system is more heavily used at some times).

If we keep all these factors except the number of fields searched (iii) and the number of hits (iv) constant we can get an idea of the relative times needed to search the different fields.

From a small sample it appears that there is no great difference in search times for the 245, 650/1 and 690 fields - though possibly the 650/1 has slightly shorter times and the 690 slightly longer. Searching all these three fields together is only one third more time-consuming than searching any one field on its own. Weighted searching is slower than boolean - a fact probably due to the way the programs are written (they were originally only set up to do boolean searches). Both LC classification number (050) and RIN (692) searches are much faster than searches on the verbal fields, but both these types of search give low performance figures.

4.2.3 Times for scanning BNB Weekly List

Most librarians took about thirty minutes per week to scan this list.

4.3 Relations with a bureau

Most of the time-sharing bureaux in London were contacted at the beginning of this project but only Cybernet Time-sharing were able to offer any previous experience in the field. They adapted their Freesearch programs, written in Fortran, at a cost of £500-the main modification made it possible to convert the MARC records into a form which could be used by the search program. We then used the programs on their time-sharing system at a cost of £150 per month for 30 hours connect time (including CPU time) and paying about £90 per month for disc storage. The Freesearch programs are available for purchase from Cybernet or for use on a bureau basis on their Sigma 8.

Our MARCAS system was cheap and quick to set up and ran fairly error free. It kept well within our cost and time estimates - the response time of the on-line system was very rarely more than a few seconds and was usually immediate (this depended on the number of people using the computer and thus tended to be longer in the afternoons

and towards the end of the week). The good service given by Cybernet and their interest in the project have amply justified our choice of bureau.

5. Use of batch system

The running costs of the system used in this project have been discussed in an earlier chapter.

Obviously batch systems must be considered for an on-going service and costed in order to provide a basis for planning a pilot service.

5.1 Other batch systems in the UK

Several systems offer MARC-based book notification [1,5]. In the UK these are basically in-house operations with a few subscribers in related institutions. For instance, the Birmingham Cooperative Libraries Mechanisation Project gave a service for some time to the Bath University Social Science Information Officer[7]. Bumpus, Haldane and Maxwell Ltd and Richard Abel and Company[9] both offer book notification services for their customers, which are in both cases free of charge - although there is a stated obligation to purchase some books from the supplier.

B, H & M Ltd's brochure says "As you are no doubt aware, it costs a great deal of money to provide this bibliographic service. We would therefore appreciate it if you make positive arrangements to order from us at least those publications that have been brought to your notice by Maxwell's bibliographic service."

5.2 Costs of other systems

It is worth examining what information is readily available concerning the costs of operating a batch system using BNB or LC MARC tapes and seeing if these can be applied in our case. Unfortunately we have not located much data expressed in computer time which could be translated to another environment. Mauerhoff gives detailed costs but does not give the times used nor the rates charged for his computer, which is an IBM S360/50 with 256K bytes of memory. Using the SELDOM programs which they developed, the cost per profile for running 82 profiles is $116 annually for searching Library of Congress tapes each week. A charge of $0.37 is made for alterations in profiles[4].

The Aldermaston AMCOS system selects items from the MARC tapes using 58 DC numbers as search keys. Selection and printing the 10% of the files which are 'matches' (about 200 items per week) costs £18 each week using an IBM S360/75, which involves 15 minutes elapsed computer time and 1.08 minutes CPU time. £6 of the £18 was for translating the tapes from the ASCII code of the MARC tapes to EBCDIC, the internal IBM code [10], £12 of the £18 is comparable to one of our library profiles.

5.3 MARCAS batch system costs

A batch version - or more correctly - a remote entry for batch processing of the MARCAS programs became available in November, 1972. By rerunning some of the profiles on the last week of our test data in batch, we arrived at computer costs of £1.30 per library for searching Library of Congress and BNB tapes.

Using the new MARCAS programs would entail small development costs in programming. One additional small program is required to reformat the "hit" records. When usage reaches 10 profiles, which hopefully it would do very quickly, some programming is required which would in fact merge the profiles for running, making this a true batch job and presumably

reducing the costs further. £200 would be required for both these alterations to the system. They should be done after a survey to determine the most desirable form of presentation.

5.4 Large systems

If more than 20 profiles, are to be processed a more elegant file structure should be investigated. Tell[11] has found, using an IBM S360/30 with 64K memory, that at about 2000 profile terms, the overhead required by this additional file handling is off-set by faster search times, and between 5000 and 10,000 terms there is virtually no increase in search time, using hashcoding and tree-structure searches. Some comparison of available systems should be made if the demand is sufficient. The Canadian SELDOM programs, the Swedish ABACUS and other programs designed for large scale SDI (e.g. more than 2000 search terms per run) could fruitfully be explored.

5.5 Provision of standard interest profiles

An organization which aims at being cost recovery would have difficulty in producing a library-oriented service which is cheaper than the Oklahoma, Florida and Canadian services, none of which is marketed in the U K and all of which are aimed at individuals or standard interests. (See for example, table 7). Likewise none of the British in-house services are being marketed.

A combined US - UK service is not available yet but undoubtedly the Canadian service and others will offer it soon, at least in North America.

It appears as if the pattern of offering a two-tier service - both specialized profiles and standard profiles - is economically a very practical mix and is the one we should explore in any future work we undertake.

Table 7 Standard Profiles for MARC

1. Oklahoma Department of Libraries

South West
Library science
Bibliography and Reference
Political science and law
Drug abuse
American Indians)
Environmental science) planned.
)

2. Florida University Library (Nov 1971)

The MARC II-IS service includes at the present time Current Aware Searches (CAS), Retrospective Searches (RS), and Standard Interest Profiles (SIP) both current and retrospective.

A number of Standard Interest Profiles for the MARC II search system are already available. New MARC II SIP's will be added as demands develop. Currently available MARC II SIP's include:

MECHANICAL ENGINEERING
TAXATION (STATE & LOCAL)
HIGHER EDUCATION
MULTI-MEDIA
LIBRARY SCIENCE
AMERICAN REVOLUTIONARY WAR PERIOD
CIVIL WAR PERIOD
AMERICAN COLONIAL PERIOD
PRIVATE AVIATION
MARKETING
SALESMANSHIP & SALES MANAGEMENT
STATE GOVERNMENT
LOCAL GOVERNMENT
JAPAN
BLACK STUDIES
CRIMINAL JUSTICE
URBAN PROBLEMS
RURAL PROBLEMS

6. Conclusions and recommendations for future work

6.1 Implications of findings about retrieval characteristics

6.1.1 Recall and precision

Based on the results given in section 4, it seems that a retrieval system intended to aid book selection by special librarians could provide a list of recently published books of which about half would be of interest to the library for which the profile had been designed. As the total number of references received would be small, on average, this low precision does not seem worrying. Recall would be better than 50%. These figures are for natural language searching of title and subject headings using boolean logic, which produced our best results. Class number searching could be used in cases where high precision is requested.

Such a service could not provide the only book selection aid for a special library's 'core' area but would provide a very useful supplement to publishers' announcements and standing orders for series known to be useful. In less central areas of interest, where complete coverage is not required, a MARC book selection service providing a list of an average of 15 books per week could save the acquisition librarian a lot of time scanning publishers' announcements, reviews, etc.

6.1.2 Currency

The information collected about the number of items retrieved which were already known to the librarian has two main implications: ⅓ of the UK books 'of importance' to them (probably 'core' books) were not known and ⅔ of the US books were not known. For books 'of interest' these figures are higher especially for items appearing on American tapes (99%) which were, by the nature of the service, even older notifications. This would indicate that a book notification system which includes the Library of Congress tapes would provide a service which libraries are not receiving

One of our users commented that European tapes would be more use than US tapes. We have no reason to doubt that we would get good currency results on these tapes since libraries get fewer publishers' announcements from abroad and currently rely more on less speedy forms of notification (i.e. reviews, advertisements in foreign journals, citations).

The (lower) precision - recall results we achieved by searching titles alone are applicable to Whitaker tapes. If these tapes were available, the currency results should be better, since these tapes contain notifications of books due to be published, as opposed to those already published and deposited for copyright as in the case of MARC tapes. Would the loss in recall be offset by the improved currency (and probably greater cost as Whitakers would probably charge a larger royalty fee)?

6.1.3 Profiling

Constructing a good natural language profile representing the entire interests of a library takes longer than constructing one for a subject specialist. With experience one should not take as long as we did. Also more interaction with the subscriber would be possible and desirable and should speed up profiling. We did not experiment with users constructing their own profiles, although they saw the profiles in the first week. If this could be done effectively costs would be reduced considerably.

Another area worth considering is automatic profile construction using LC Subject Headings and other indexing terms. By analysing the terms used to index books selected as relevant over a period of 2 to 3 months, a profile could be constructed. Any additional subject headings used on relevant books retrieved could be analyzed by the user to

see if it were a heading of interest to them.

Until methods of reducing profiling costs are tested, our costings must be based on those found in our study. Use of known relevant items seems the most useful profiling tool. Lack of subject expertise on the part of the profile constructor was, not surprisingly, found to be a disadvantage.

6.2 Standard profiles

The main advantage of standard profiles is that the same costs are shared by a number of institutional users. Many tape-based retrieval services are marketing standard profiles, including at least two LC-MARC services (Oklahoma and Florida).

A mix of standard and specialized profiles would enable both services to be produced at a lower cost since fixed costs would be shared over a larger number of users.

Most tape-based SDI services offer some sort of standard profiles at a reduced rate. This is usually an afterthought, resulting from the file being available, lower demand than anticipated for individual profiles and to an awareness that certain topics are amendable to standard profile techniques or meet a demand.

The following factors about standard profiles need to be determined for MARC tapes before a service can be offered:

1. Can a good cheap standard profile be defined as a topic broad enough to be of wide interest yet specific enough to be defined by relatively few terms? Define a strategy for locating a good profile.
2. What is the cost of processing a standard profile?
3. How many standard profiles would be required to cover a library's interest (and vice versa how much of a

library's interest can be covered by a profile)? Could a combination of standard profiles replace a specialised profile?

4. How would the provision of standard profiles affect the cost of a wider book notification service?
5. Would the user of standard profiles be the same as for the specialised profiles?
6. What standard profiles are offered by other non-MARC services?

6.3 Survey of book selection methods and needs

The small sample on which we worked can give us fairly reliable operational data on which to base the design of a larger system. However we still know very little about the demand and need for a book notification service. Before a larger experimental system is developed a survey should be carried out in an attempt to answer the following questions about present book selection practices and deficiencies:

1. How much time is spent in special libraries on book selection by what level of staff?
2. What book selection aids do they use? Which do they pay for?
3. Are they interested only in British books? - English language books?
4. Do they wish to scan a narrow range of subjects or a broad one?
5. Do they buy from just a few publishers?
6. Would they prefer
   high recall or high precision?
   expensive individualized profiles or
   cheap standard profiles?
   (including price they would be willing to pay)
7. Do they think they have a problem?
8. Do they prefer card (more expensive) or listings on A4 paper?
9. What information other than title and author should be provided:

ISBN
DC number
LC classification number
Publisher, Imprint
Corporate author
LC Subject headings
Others?

6.4 Other recommendations for future work

If there is a need for either specialized profiles and/or standard profiles, then some development work and assigning of responsibility is required.

6.4.1 Investigation of other systems for improved performance and availability

The flexibility of the MARCAS system leaves little to be desired from the user's point of view: the method of entering searches is simple and straightforward and any field can be searched. However no claims can be made about its efficiency. Although we have no data for comparable BNB or LC MARC tapes, our computer costs may be higher than those of systems searching other tape files. For this reason, trial runs should be carried out using other available software such as the SELDOM, ABACUS, and AMCOS programs and perhaps trying more generalized packages. This should begin as soon as possible but can and should extend until all the alternatives have been tested. This will not require much man-power per test and the MARCAS system can be used until a demonstrably better system has been accepted.

6.4.2 Pilot Service to Fifty users for a year at a subsidized price of £25. Participants will be sought from those responsible for book selection among industrial, government and academic organizations with various sized book purchase budgets. A variety of output formats will be used including cards, A4 listings,

various fields of information printed, etc. User reaction to the service will be monitored by questionnaire, interview, the right to cancel the service or to alter their profile.

Initially the computer processing will be carried out using the MARCAS programs in batch processing mode. Cybernet Time-Sharing Sigma 9 will be used in the first instance.

Although a service offered to 50 users at £25 a year will have an income of £1250, a further £10,000 will be needed to cover the costs of this part of the project. Manpower allocations are a problem since the major part of the profile construction work obviously must be carried out early in the project.

A charge to the user is felt to be desirable in order to insure that the participants have a real need for the service and will make real demands on it. Also such a charge will effectively eliminate individuals using a system designed for what amounts to group profiling.

From offering such a service, we hope to gain further understanding of the problems involved in running a MARC-based book notification system. Experience in profile maintenance, processing and distributing will be gained. A pilot service should give us a better understanding of the viability of such a service and the form it should take.

6.4.3 Profile development and evaluation

Responsibility for continuing customer liaison should lie with the person developing the profiles. Since fringe areas of interest shift more than core areas, ample time should be allowed for profile modification and alteration. Continuous evaluation

and customer feedback will only be possible if the manpower is made available from the start. If at all possible, the people involved in constructing profiles during the experimental period should continue to use on-line techniques for profile construction. There is no doubt that this will result in faster accumulation of "operational" profiles.

References

1. WAINWRIGHT, J. BNB MARC users in the UK;a survey,Program, 6 (4). 1972, pp.271-283.

2. STUDER, W.J. Computer-based selective dissemination of information (SDI) service for faculty using Library of Congress MAchine-Readable Catalog (MARC) records. Indiana University. 1968. Ph.D.thesis.

3. BIERMAN, K.J. and BLUE, B.J. A MARC-based SDI service. Journal of Library Automation, 3(4). December 1970, pp.304-319.

4. MAUERHOFF, G.R. A MARC II-based program for retrieval and dissemination. Journal of Library Automation, 4(3). September 1971, pp.141-158.

5. PATRINOSTRO, F.S. A survey of automated activities in the libraries of the United States. Vol.1.LARC Assoc.1971.

6. ATHERTON, P. and MILLER, K.B. LC/MARC on MOLDS; an experiment in computer based, interactive bibliographic storage, search, retrieval, and processing. Journal of Library Automation, 3.June 1970, pp142-165.

7. LINE, M.B., CUNNINGHAM, D. and EVANS, S. Experimental information service in the social sciences 1969-1971. Final report. Bath University Library. January 1972.

8. DATTA, S. and ROBERTSON, S.E. Analysis of on-line searching costs: an experiment using a commercially available reference retrieval system (SCISEARCH). Aslib Research and Development Department. February 1972.

9. CHVATAL, D.P. and OLSON, G.L. A computer-based acquisition system for libraries. In: Proceedings of the ASIS 34th Annual Meeting, Denver, November 1971. Volume 8. Greenwood Publishing Company. 1971. pp.217-226.

10. CORBETT, L. and GERMAN, J. AMCOS project stage 2: a computer aided integrated system using BNB MARC literature tapes. Program,6(1), 1972. pp.1-35.

11. TELL, B.V., LARSSON, R. and LINDH,R. Information retrieval with the ABACUS program. IAEA Symposium on Handling of Nuclear Information, Vienna. 1970.pp.183-199.

12. KING, M., HOLLOWS, A. and BAKER, R.F. Index to standard profiles available in June 1972. Birmingham University Main Library, Science Information Office. July 1972.

13. LEWIS, P.R. Systematic library use of British National Bibliography services and data. A survey of British practice. Journal of Librarianship, 2 (4) 1970. pp.211-226.

Appendix II

Potential for use with other MARC format tapes

There are certain limitations connected with on-line information retrieval systems: they are expensive; it is not possible, usually, to have large data bases available. Thus their main use would appear to be for profile construction and also for demonstrations. However the MARCAS program works well and is available for use on any MARC II format tape with only slight modifications to the initial program (i.e. removal of the juvenile indicator and English language tests). Enquiries have already been made concerning their use from the Road Research Laboratory and the National Council of Educational Technology.

The following tapes are available or will shortly be made available, in MARC II or compatible format (B.S.4748.1971)

Bibliography of Agriculture

CAB

COMPENDEX

Current Index to Conference Papers in Chemistry, Engineering, Life Sciences

ERICTAPES

INSPEC

PANDEX

PAIS - Psychological Abstracts

TAB Indexes

UNESCO